UDC 621.391 DOI: 10.20535/2411-2976.12025.59-66

VOIP SYSTEM WITH HIGH AVAILABILITY

Serhii O. Pryimak, Serhii O. Kravchuk

Educational and Research Institute of Telecommunication Systems
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine

Background. With the exponential growth of the internet and an increasing number of VoIP (Voice over IP) deployments arose a need to manage and scale the systems in terms of high availability and minimal downtime.

Objective. The purpose of the paper is to reveal methods of increasing the VoIP system's availability. Main challenges associated with the operation of VoIP systems are analysed, particularly the need to ensure service continuity, fault tolerance, scalability and resource efficiency.

Methods. Increasing the availability of a VoIP system is based on the use of free software: Kamailio (SIP signalling server), RTPengine (media traffic processing), Redis (in-memory database for storing RTP transactions), and Linux Ubuntu as a host operating system.

Results. The implemented architecture reflects automatic failover of signalling servers and media gateways, load balancing and seamless switching to a backup node in case of failure of one or more components.

Conclusions. The approach given drastically differs from others, where failure of one of the key elements (the signalling server or the media gateway) leads to immediate termination of all active connections. An experimental simulation of the system was performed by deploying a high availability schema depicted in Fig. 2 in a virtual environment and confirmed its high efficiency: even in the event of 50% of component failure, the system remains operational and does not interrupt any call.

Keywords*: VoIP system; SIP; RTP; high availability; reliability.*

1. Introduction

Permanent demand for quality voice traffic on packet networks is driving high interest of developers to improve technology of voice over the Internet protocol VoIP (Voice over IP) [1, 2]. The technology given is not based on the traditional public switched telephone network PSTN; calls are made by transferring data over the Internet protocol.

At the moment, there are certain limitations in a VoIP telephony architecture that are oriented to provide high-quality telephone services, especially in a highly loaded system. It is extremely important to ensure a stable work in organizations where customers are actively using VoIP services [3, 4]. Such restrictions can lead to the fact that the organizations, either to improve service delivery or to avoid interruptions in work, have to deploy custom solutions that often require significant financial investments. Main difficulties faced by organizations using open source software solutions for telephony services are a secure and stable system work in conditions of high parallelism with the presence of huge number of requests.

Today, a lot of interest is shown to ongoing research related to the further development of VoIP systems.

1. High availability: VoIP systems are responsible for providing uninterrupted access to communication in the event of failures [5, 6]. This can be achieved by different methods: redundancy, server clustering, and multiple routes for data transmission.

2. Quality of Service (QoS) requirements: one of the main problems of VoIP system availability is the provision of stable voice transmission capacity, which includes low latency, minimal packet loss and high availability [7]. This is achieved by proper distribution of resources, traffic planning and adaptation to network changes.

3. Reliability and resiliency: VoIP systems can be protected from various types of failures, including server failures, security failures and DDoS (distributed denial-of-service) attacks. Research in this area is aimed at developing mechanisms for rapid renewal after failures.

4. Security: since VoIP systems are functioning via the Internet, they become exposed to attacks such as IP address spoofing, call interception, attacks on registration servers (SIP (Session Initiation Protocol) servers) and DDoS attacks. Enhanced accessibility often includes protection against these risks by using data encryption, user authentication and other security methods.

5. Scalability: with the growing number of customers and traffic, VoIP systems are required to become stable and ensure high-capacity communications without interruption, thus, researches also span architecture development that can be scaled

according to needs.

6. Efficiency of resource utilization: optimization of resource utilization and computation are key values in the scope of high availability. Modern research is exploring a variety of methods for compressing voice data, managing data throughput, and minimizing power consumption.

As more and more companies move to IP telephony, the demand for secure, uninterrupted operation of VoIP systems is growing. Therefore, the relevance of research in the field of increasing availability of VoIP systems is growing due to the wide expansion of such systems both in the private and commercial environments. Some other factors which make this research field relevant are as follows:

- widely spread remote work: due to the COVID-19 (SARS-CoV-2) pandemic and the growing popularity of remote work, VoIP technologies have become key communication tools;

- cost savings: companies are actively switching to VoIP due to the opportunity to reduce telephone costs;

- business globalization: the need for high-quality international communication is growing, which requires reliable and affordable VoIP solutions;

- mobility and flexibility: VoIP allows you to integrate telephone systems with other communication services (video conferencing, instant messengers), which increases the flexibility of communications.

In general, the availability of VoIP systems means ensuring the constant and uninterrupted operation of these systems, even in the event of problems with the network, equipment or under conditions of increased load. Currently, further development of VoIP systems by increasing availability is associated with the following areas:

- Cloud integration: Cloud-based VoIP solutions have become the foundation for many business systems because they provide availability through decentralized architectures. Current research covers methods for ensuring availability over distributed networks with cloud infrastructure;

- Communication protocol optimization: SIP [1] and RTP (Real-time Transport Protocol) [8] play a key role in ensuring the availability and quality of VoIP connections. Research is aimed at optimizing these protocols to reduce latency and increase reliability;

- DDoS protection: Although VoIP systems remain vulnerable to DDoS attacks, modern solutions include traffic filtering methods, intelligent network solutions, and cloud-based defence mechanisms;

- Backup and load balancing mechanisms: Methods implemented allow automatic traffic redirection in the event of server failure or overloads.

The following issues arise and need to be solved:

1. Ensuring quality over unstable Internet connections: even using high-speed Internet, VoIP communication can suffer from packet loss, latency, and fluctuations in network parameters.

2. Integration with different network technologies: The need to ensure compatibility of VoIP systems with different types of networks and devices remains a challenge.

3. Protecting privacy and ensuring security: Despite the presence of encryption and security protocols, VoIP systems remain a target for hackers.

4. Effective scaling: With the growing number of users, VoIP systems have to adapt quickly, which requires the development of new approaches to scaling.

The goal of this research is to develop and improve methods for increasing VoIP systems' availability by:

a) creating adaptive traffic management systems that provide a balance between availability and quality of service;
b) conducting experiments on scaling VoIP systems to verify their reliability under increasing load;
c) developing new approaches to ensure VoIP networks security.

These areas are based on modern scientific research in the field of telecommunications and cybersecurity, which confirm the importance and relevance of these areas of development.

An example of building a VoIP system with real-time encryption of voice and signal traffic is considered in [9], in [10] main focus is on building a secure computer network for a VoIP system and typical load indicators of VoIP servers are given, but none of these papers considering aspects of fault tolerance of software and hardware elements of a VoIP system. In [11], several methods of increasing the availability of a VoIP system were considered, but none performed takeover of existing RTP voice streams in the event of failure. In the case of implementing the system proposed by the authors, there is no impact on services even if 50% of its components fail.

Of interest for improving VoIP systems are approaches based on free software Kamailio, RTPengine and Redis [17]. But their application is not sufficiently revealed in the available papers.

Therefore, the purpose of this research is to develop a method for increasing the availability of a VoIP system created on the basis of free software Kamailio, RTPengine and Redis without affecting the service if 50% of the system elements fail.

2. Traditional VoIP system

Nowadays, VoIP systems are widely used in mobile and fixed networks for traffic handling [12]. The traditional topology of a VoIP system is shown in Fig.1 [13].

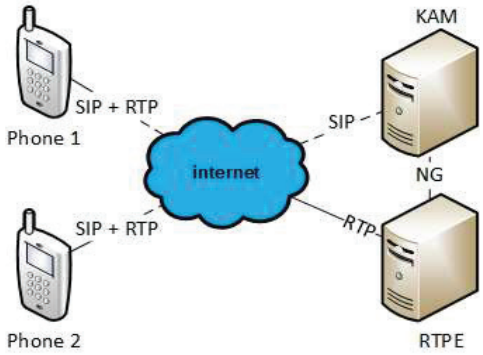

Fig. 1. Traditional VoIP system

The traditional VoIP system reflected in Fig.1 consists of a SIP signalling server (KAM) Media gateway (RTPE), which are responsible for SIP signalling and voice RTP packets handling, respectively and SIP phones Phone 1 and Phone 2. It is impossible to increase the given system's availability, and in case of failure of one or more components, we observe signalling and/or voice traffic interruption. Typically, such systems use software that combines signalling and media servers, such as Asterisk, FreeSWITCH, on the Linux operating system. In some cases, a virtual environment such as VMware, VirtualBox, HyperV, Docker, or others is used.

3. Some aspects of Session Initiation Protocol (SIP)

SIP is a text- and request-response-based protocol that is similar to the Hypertext Transfer Protocol (HTTP) in design. It has been developed by the Internet Engineering Task Force (IETF) and gained popularity in recent years, mostly in Voice over IP (VoIP) applications and became a standard in 1999.

The main signalling functions of the protocol are as follows:

- Location of an endpoint;
- Contacting an endpoint to determine willingness to establish a session;
- Exchange of media information to allow a session to be established.

SIP is a key protocol used for media transferring over an IP network, thus, SIP is a simplified version of signalling protocol widely used in IP telephony, where services can be controlled by the subscriber.

SIP network uses a signalling server for its operation, which is a device that transmits or manages signalling information [14]. Signalling servers can act as a proxy server [15]. A proxy server represents the interests of the user in the network. It accepts requests, processes them, and performs appropriate actions. A proxy server consists of a client and a server part, so it can accept calls, initiate requests, and return responses.

There are two types of proxy servers:

- with state preservation (stateful). Such a server stores received requests and related newly generated requests in its memory until the transaction is completed.
- without state preservation (stateless). Such a server simply processes received requests. But it is impossible to implement complex, intelligent services on its basis.

Another element of the SIP network is a media gateway. This is a device that provides interaction between different types of networks for the transmission of multimedia content, such as voice, video, or text [15]. A media gateway can allow voice messages to be exchanged between IP networks (VoIP) and traditional telephone networks (PSTN) [16].

SIP is a simple signalling protocol for establishing, modifying, and terminating voice and multimedia connections in VoIP and multimedia conferencing sessions. It is a client-server protocol and is similar to the HyperText Transfer Protocol (HTTP) in both syntactic and semantic terms. It has text-based requests and responses that contain header fields that convey information about the service and characteristics of the connection.

As with HTTP, the client communicates with the server using so-called request methods. SIP defines 6 basic methods - INVITE, ACK, OPTIONS, BYE, CANCEL, and REGISTER, as well as various response codes, divided into 6 classes (Table 1).

Table 1. SIP response codes

Class	Value
1xx	Informational
2xx	Success
3xx	Redirection
4xx	Request failure
5xx	Server failure
6xx	Global failure

Basic SIP methods:

INVITE – Initiates a call or session by inviting a

user to participate in a communication session.

ACK – Confirms the receipt of a final response (2xx) to an INVITE request.

OPTIONS – Queries the capabilities of a SIP server or endpoint without establishing a session.

BYE – Terminates an ongoing session between two users.

CANCEL – Cancels a pending INVITE request before the session is established.

REGISTER – Registers a user's location with a SIP server to receive calls.

Apart from the basic, there are extended methods:

UPDATE – Modifies an existing session without affecting the dialogue state (e.g., changing media parameters).

INFO – Sends mid-session signalling information (e.g., DTMF tones, application data) without affecting the session state.

PRACK (Provisional Acknowledgement) – Acknowledges reliable provisional responses (1xx) in SIP to ensure message reliability.

SIP is independent of the model and scale of communication or conferencing and of the packet layer, requiring only datagram delivery services without acknowledgement, since the reliability of their delivery is ensured by its own mechanism [15].

4. System availability assessment

General system availability can be calculated according to the formula below:

$$A = \frac{MTBF}{MTBF + MTTR}$$

where: MTBF - Mean Time Between Failures; MTTR - Mean Time to Repair.

We assume the following data for the calculation:

– MTBF for traditional and highly available systems is equal to two weeks → 336 hours →1209600 sec.
– MTTR for a traditional system is 60 min. → 3600 sec. This is the amount of time to restore the server from backup in case of failure;
– MTTR for a highly available system is 1 sec, which is the time to perform handover to the standby server.

Thus, for a traditional VoIP system, Availability is:

$$A = 1209600 / (1209600+3600) = 99{,}7\%,$$

while for a highly available VoIP system, we have the following Availability:

$$A = 1209600 / (1209600+1) = 99{,}9999\%.$$

5. Highly available VoIP system

A highly available VoIP system is a voice communication system that uses IP packets to transmit voice (video, etc.) and is designed to ensure continuous operation, even in the event of failure of individual components, software and hardware or network problems. The main goal of such a system is to minimize downtime, ensure fault tolerance, perform load balancing, as well as duplication and resource redundancy [14].

The scheme of a highly available VoIP system is presented in Fig. 2.

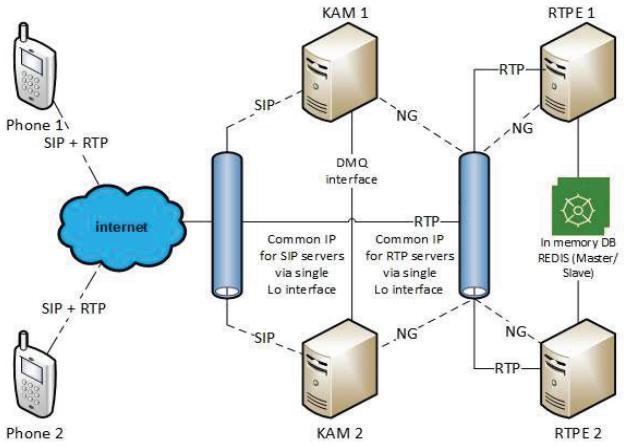

Fig. 2. Highly available VoIP system

The system given consists of the following components:

- SIP signalling servers KAM1, KAM2 for SIP traffic handling;
- Media Gateways RTPE1, RTPE2 for voice RTP packets handling;
- In-memory DB Redis for storing RTP transactions parameters;
- Subscriber SIP terminals Phone 1, Phone 2.

We are using open-source software products:

- Kamailio [15] for SIP server;
- RTPengine [15] for media gateway;
- REDIS [17] as RTP transaction data storage;
- Keepalived to establish handover between nodes in case of failure.

Redundancy is organized in the following way: between KAM1 and KAM2 servers, we have a logical DMQ (data message queue) [15] interface [15] to facilitate SIP transactions (SIP call-id) [1] propagation

and replication. RTPE1 and RTPE2 servers store active RTP transactions into independent databases inside the Redis cluster by using the pub/sub protocol while having access to each other. Redis itself is organized as an active/backup cluster with automatic replication. Each Pair of KAMx, RTPEx and REDISx has a personal virtual IP address assigned automatically on an active node, which makes transparent access to any active node of the system. Switchover time to hot standby is 1 sec.

6. Typical call scenarios

Let's consider voice call processing scenarios: as already known, a voice call in a VoIP network consists of several components:

- signalling stream (in this case, SIP);
- voice data stream (RTP).

In the scheme shown in Fig. 2, the following priorities are set by default: the signalling server KAM1 manages media gateways RTPE1 (main) and RTPE2 (backup), and KAM2 manages media gateways RTPE2 (main) and RTPE1 (backup). Both media gateways use the REDIS database to store RTP transaction parameters. REDIS, in its turn, monitors activity between RTPengine instances and performs emergency sessions switchover from one instance to another.

In this implementation, only one KAM + RTPE pair can be active and process voice calls, another pair is in hot standby mode.

In the case of all system functioning correctly without issues, the voice call is processed as follows: signalling is processed by KAM1 and at the same time, it controls RTPE1, responsible for the voice RTP stream between subscribers Phone1 and Phone2. If the call is terminated by one of the subscribers, KAM1 sends a command to RTPE1 to terminate the RTP stream, and then the call is considered finished.

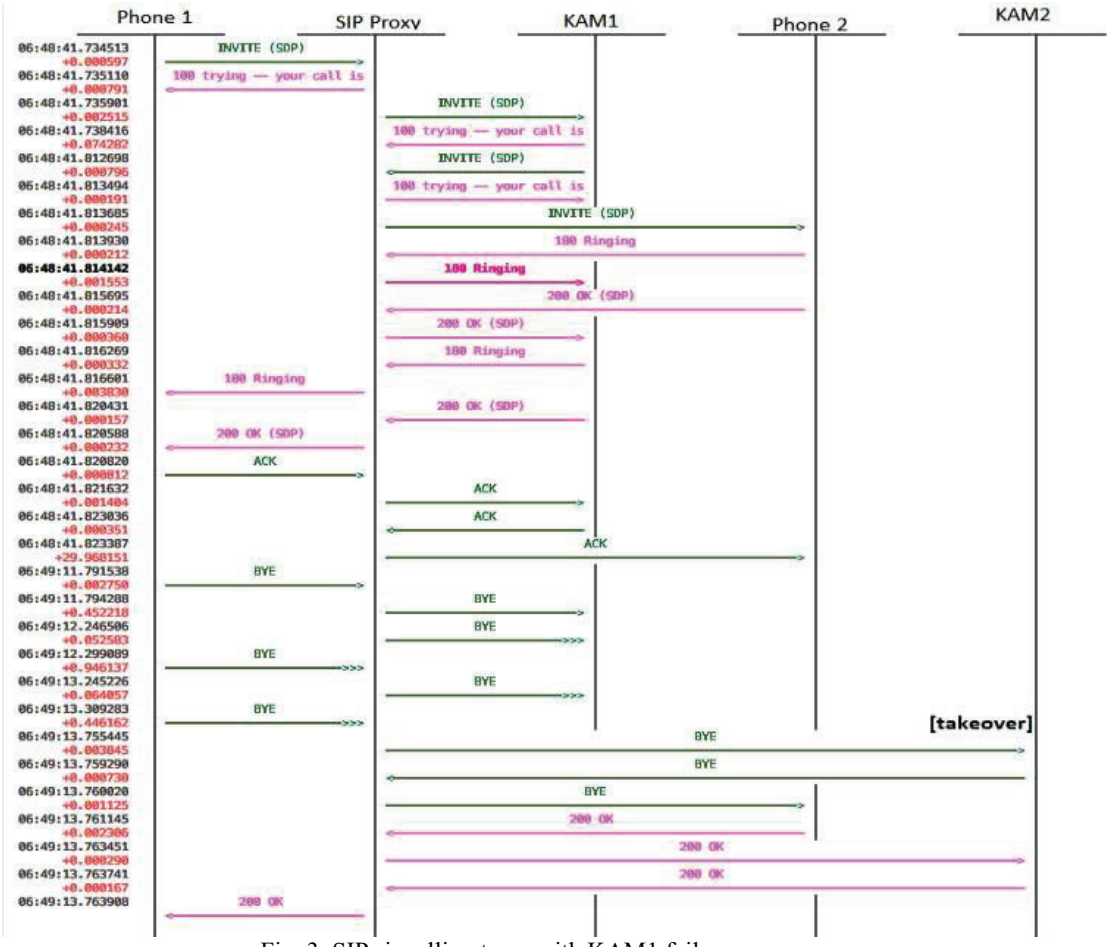

Fig. 3. SIP signalling trace with KAM1 failure

Let's consider cases when we have a failure of one or more system elements:

1. The connection was successfully established according to the algorithm described above, but later we had a total outage of the KAM1 signalling server. Voice stream continues functioning, but such a call would “never” be completed in the absence of KAM2, which controls the SIP transactions of the KAM1 server via the DMQ interface. Thus, KAM2 takes over all active signalling transactions of KAM1.

2. The connection was successfully established according to the algorithm described above, but later we had a failure of RTPE1. In this case, thanks to the REDIS database, all RTP transactions from faulty RTPE1 can be taken over by RTPE2 while KAM1 continues to process signalling.

The object of research is to conduct two experiments:

A. Preservation of a voice call and its correct processing after a software or hardware failure of the KAM1 signalling server managing this call.

B. Preservation of a voice call and its correct processing after a software or hardware failure of the RTPE1 media gateway processing the voice stream.

The essence of experiment A is as follows:

1) subscriber Phone1 makes a call to subscriber Phone2 in the system shown in Fig. 2;

2) having established a connection between these subscribers, we artificially invoke a failure in the KAM1 signalling server by turning off the virtual computer;

3) since all SIP transactions are “mirrored” in the KAM2 signalling server, and Keepalive software implements the correct status of the signalling server (in this case, KAM2 changes its status from backup to primary) and continues managing all transactions previously handled by KAM1;

4) if the KAM2 server receives signalling messages in the scope of a SIP transaction that had been created by KAM1, KAM2 looks for this transaction in its database and continues call processing.

The signalling trace of such a voice call described above is shown in Fig. 3. The moment of the KAM2 signalling server response is marked with the [takeover] label.

The essence of experiment B is as follows:

1) subscriber Phone1 makes a call to subscriber Phone2 in the system shown in Fig. 2;

2) SIP transaction is processed by the KAM1 signalling server, and the voice stream is formed by the RTPE1 media gateway;

3) once the connection between these subscribers is successfully established, we artificially invoke a failure on RTPE1 by turning off the virtual computer;

4) due to the keepalived software, the RTPE2 media gateway becomes active and retrieves RTP transitions of faulty RTPE2 from the REDIS database with further processing.

Fig. 3 reflecting a voice call between subscribers Phone1 and Phone2 via a highly available VoIP system, which consists of subscriber SIP terminals Phone1 and Phone2, signalling servers KAM1 and KAM2 and SIP proxy, which in this case acts as a SIP message relay and provides the ability to collect a full trace of the signalling exchange between the above-mentioned system elements. Let us consider in more detail how the voice call processing.

At 06:48:41.734513, a SIP INVITE message is received from subscriber Phone1 purpose of which is to create a new voice session with a unique SIP call-id [1].

As a response to the INVITE, subscriber Phone1 receives a 100 Trying message (request in processing) from SIP Proxy, which, in turn, relays the INVITE message to the active by default signalling server KAM1.

KAM1 checks the internal routing table, finds the IP address of the subscriber Phone2, changes it in the "To:" field of the INVITE message, and forwards it to the SIP Proxy at time 06:48:41.812698.

Having received the INVITE, the SIP Proxy forwards it to the end subscriber Phone2 and, as a response, receives a 180 Ringing message (calling the subscriber), which is broadcast to the signalling server KAM1.

At 06:48:41.815692, a 200 OK message (subscriber answered the call) was received from the subscriber Phone2 by the SIP Proxy, broadcasting it to the signalling server KAM1 and then to the end subscriber Phone1.

Having received a 200 OK message that the subscriber accepted the voice call, Phone1 sends a confirmation in the form of an ACK message, which in turn is broadcast to the end subscriber Phone2 via SIP Proxy and KAM1 (time interval 06:48:41.820820 - 06:48:41.823387).

We invoke artificial failure of the KAM1 signalling server (turn off the virtual machine).

At 06:49:11.791538, the subscriber sends a BYE call completion message to the SIP Proxy, which then broadcasts to the KAM1 signalling server.

Having received no response from the KAM1 signalling server, the SIP Proxy sends a BYE message to the KAM2 server, which identifies the transaction by the SIP call-id previously received via the DMQ interface from the KAM1 server and forwards it to the end subscriber Phone2.

Subscriber Phone2 sends a call completion confirmation message 200 OK to subscriber Phone1 via chain Phone2 → SIP Proxy → KAM2 → SIP Proxy → Phone1.

Thus, we see that after the failure of the signalling server KAM1 SIP dialogue does not break and at 06:49:13.755445 it processed by the signalling server KAM2. Switching to hot standby is performed automatically so there is no impact on service.

7. Conclusion

A structural model of a highly available VoIP system, shown in Fig. 2, has been developed, and its simulation performed using software for virtualization of computer systems. Availability of the system simulated is 99.9999%, and even in the event of a failure of 50% of system components, all existing and new voice calls continue to be processed.

A method for increasing the availability of a VoIP system in the event of a failure of one or more of its components is presented. This method differs from traditional approaches by preserving all existing media calls in the event of a failure of one or more components using free software. The main difference of this method is as follows: if one of the two servers (KAMx and RTPEx) fails, the system does not lose any established calls and is able to process new ones. The system has a single-entry point (IP addresses for receiving/sending calls from/to the outside), which is unchanged in the event of a failure of system elements.

Typical failure scenarios of the KAM1 signal server and RTPE1 media gateway components are presented.

References

1. Rosenberg J., Schulzrinne H., Camarillo G., Johnston A., Peterson J., Sparks R., Handley M., Schooler E. SIP: Session Initiation Protocol. RFC 3261. Internet Engineering Task Force (IETF) 2002, 269 p.

2. Alo U.R., Nweke H.F., Voice over Internet Protocol (VOIP): Overview, Direction and Challenges. Journal of Information Engineering and Applications, Vol.3, No.4, pp. 18-28, 2013, Retrieved from: https://www.researchgate.net/publication /342_Voice_over_Internet_Protocol_VOIP_Overview_Direction_And_Challenges

3. Vetoshko I.P., Kravchuk S.O. Possibilities of improving the voice services quality in 5G networks. Information and Telecommunication Sciences. 2023. Vol.14, No 2. pp. 9-16, https://doi.org/10.20535/2411-2976.22023.9-16

4. Ahson M. VoIP Handbook: Applications, Technologies, Reliability, and Security. Ilyas, CRC Press, 2009. 472 p.

5. Habraken J. Practical Cisco Routers. QUE USA, 1999. 315 p.

6. Roy O.P., Kumar V. A Survey on Voice over Internet Protocol (VoIP) Reliability Research. IOP Conference Series: Materials Science and Engineering. 6th International Conference & Mathematical Sciences (ICCM 2020) Vol. 1020, 2020. pp. 12-19 Nirjuli, India, DOI 10.1088/1757-899X/1020/1/012015.

7. Snir Y., Ramberg Y., Strassner J., Cohen R., Policy Quality of Service (QoS) Information Model, RFC 3644. Internet Engineering Task Force (IETF), 2003. 73 p.

8. Schulzrinne H., Casner S., Frederick R. RTP: A Transport Protocol for Real-Time Applications. RFC 3550. Internet Engineering Task Force (IETF), 2003. 104 p.

9. Tulemu W., Design of an asterisk-based VoIP system and the implementation of security solution across the VoIP network. World Journal of Advanced Research and Reviews, 2024, Vol.1, № 23(01), pp. 875–906 DOI: https://doi.org/10.30574/wjarr.2024.23.

10. Tobiloba K.A., Building a Secure Asterisk-Based VoIP System Design and Implementation, Researchgate, 2023. Vol. 4, № 9(82). pp. 4–11. Retrieved from: https://www.researchgate.net/publication/388707610.

11. Martin A., Gamess E., Urribarri D., Gómez J. A Proposal for A High Availability Architecture for VoIP Telephone Systems based on Open Source Software. (IJACSA) International Journal of Advanced Computer Science and Applications, 2021. Vol. 9, No. 9, 20. 2023. Retrieved from: https://thesai.org/Downloads/Volume9No9/Paper_1-A_Proposal_for_a_High_Availability_Architecture.pdf

12. The Sipwise C5 CE Handbook mr8.5.10. 400 p. Retrieved from: https://www.sipwise.com/doc/mr8.5.10/handbook-ce.pdf.

13. Wu W. Packet Forwarding Technologies. Auerb Publications. 2008. 446 p.

14. Ahmed A., Madani H., Siddiqui T., Voip Performance Management and Optimization, Cisco Press. 2010. 448 p.

15. Mierla D.C., Modroiu E.R. “SIP Routing with Kamailio” 2022 356 p.

16. Pal S., Gadde R., Latchman H.A. On The Reliability of Voice Over IP Telephony. Computer Science, Engineering. 2011. Retrieved from: https://www.iiis.org/CDs2011/CD2011IMC/CCCT_2011/PapersPdf/TA224EV.pdf.

17. Carlson J.L. Redis in action. 2013. 320 p.

Приймак С.О., Кравчук С.О.

Високодоступна VoIP система

Навчально-науковий інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, м. Київ, Україна

Проблематика. З експоненційним зростанням Інтернету та збільшенням кількості VoIP систем виникла потреба в управлінні та масштабуванні систем з точки зору високої доступності та мінімального простою.

Мета дослідження. Ця стаття розкриває методи підвищення доступності системи VoIP (голос через IP). Проаналізовано основні виклики, пов’язані з функціонуванням систем VoIP, зокрема необхідність забезпечення безперервності обслуговування, відмовостійкості, масштабованості та ефективності використання ресурсів.

Методика реалізації. Підвищення доступності системи VoIP базується на використанні безкоштовного програмного забезпечення: Kamailio (сервер сигналізації SIP), RTPengine (обробка медіа-трафіку), Redis (база даних у пам’яті для зберігання транзакцій RTP) і Linux Ubuntu як хост-операційна система.

Результати досліджень. Реалізована архітектура відображає автоматичне перемикання серверів сигналізації та

медіа-шлюзів, балансування навантаження та плавне перемикання на резервний вузол у разі відмови одного або кількох компонентів.

Висновки. Даний підхід кардинально відрізняється від інших, де збій одного з ключових елементів приз-водить до негайного припинення всіх активних з'єднань. Було виконано експериментальне моделювання системи шляхом розгортання схеми високої доступності, зображеної на рис. 2 у віртуальному середовищі, яка підтвердила свою високу ефективність: у разі виходу з ладу 50% компонентів система залишається працездатною та не перериває жодного виклику.

Ключові слова*: система VoIP; SIP; RTP; висока доступність; надійність.*

Received by the Editorial Office
March 21, 2025

Accepted for publication
June 9, 2025